Theoretical aspects of Neural Network application for signal/background separation

Dmitri Smirnov (smirnov@unm.edu)
Physics & Astronomy Department, University of New Mexico

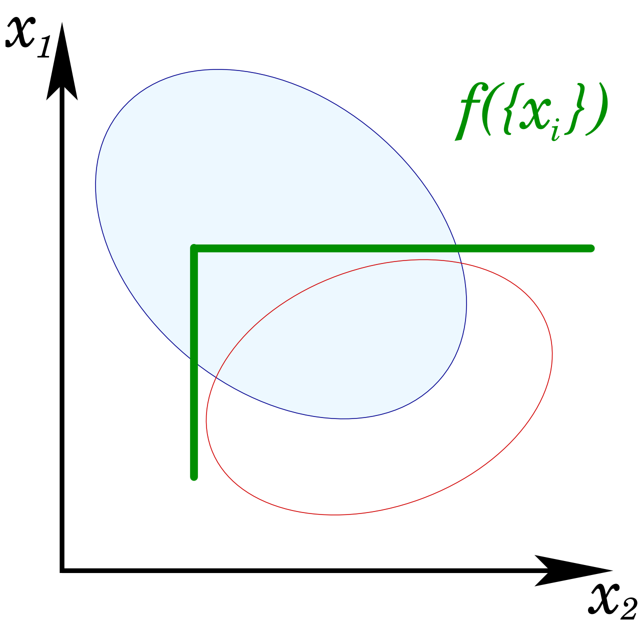
1 June 2002

Outline

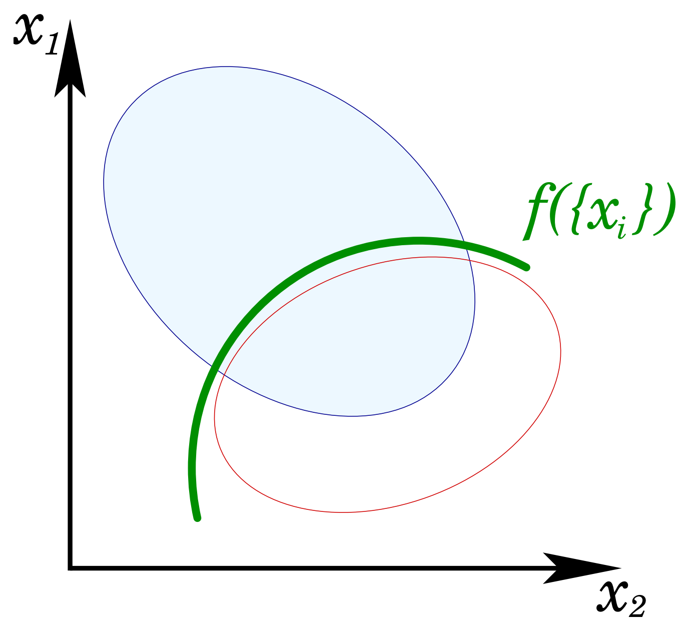
- ✓ Signal/background separation. Why ANN?
- ✓ Kolmogorov theorem $\Rrightarrow$ one output + one hidden layer.
- ✓ Optimization of input variables.
- ✓ Estimations for number of hidden units.
- ✓ Conclusion.

Why Artificial Neural Network?

- Event information can be coded by several kinematic variables $\{x_i\}$
- Ideally final answer has 2 states (1 bit): **signal** (1) and **background** (0)



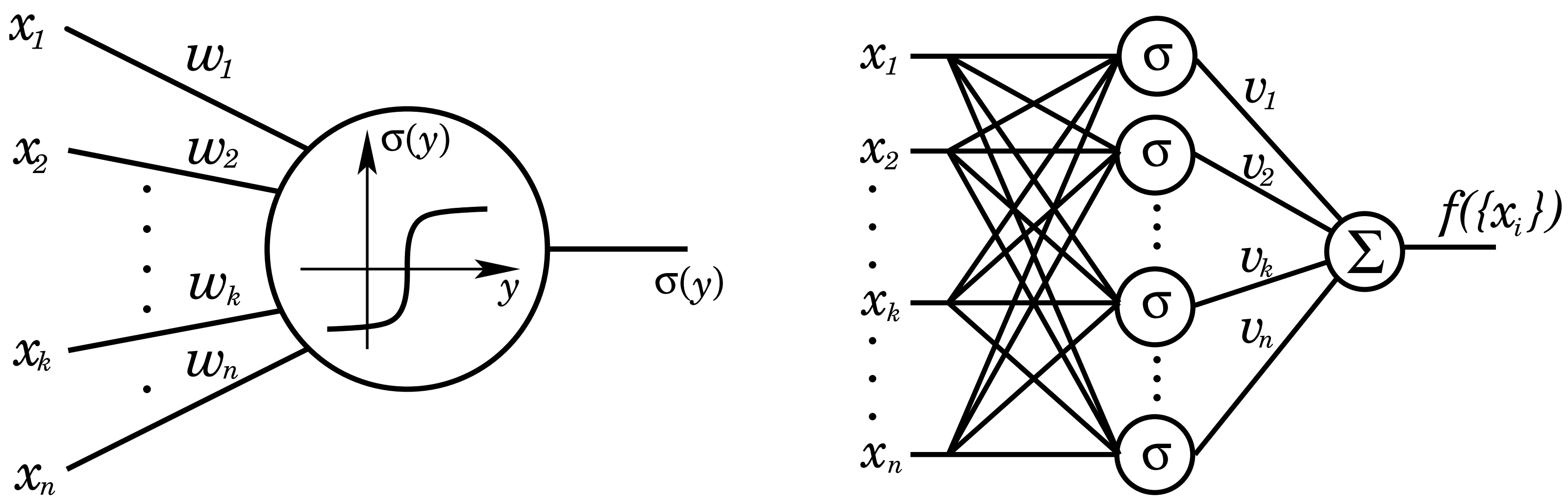
- Classic cuts are essentially linear and "rectangular"



- NN "cuts" depend on input values of all variables at the same time

ANN with one hidden layer and one output

♠ H is the number of hidden units, w_{ij} and v_j are weights, x_i are inputs, $\sigma(y) = \frac{1}{1+e^{-\alpha y}}$ is sigmoid function.



$$f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{H} v_j \sigma(w_{1j}x_1 + w_{2j}x_2 + \ldots + w_{nj}x_n) \,.$$

Theoretical representation of $f(\{x_i\})$?

♠ Kolmogorov theorem: any continuous function of n variables $F(x_1, x_2, \ldots, x_n)$ can be presented in the form of

$$F(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{2n+1} g_j \left\{ \sum_{i=1}^{n} h_{ij}(x_i) \right\},$$

where g_j and h_{ij} are any continuous functions, h_{ij} does not depend on function F.

♠ Weakening of Kolmogorov theorem's conditions. $\forall\ \varepsilon > 0\ \exists\ H$, $\exists\ \{w_{ij}\}$, and $\exists\ \{v_j\}$, such as

$$|f(x_1, x_2, \ldots, x_n) - F(x_1, x_2, \ldots, x_n)| < \varepsilon\,,$$

$$\boxed{f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{H} v_j \sigma(w_{1j}x_1 + w_{2j}x_2 + \ldots + w_{nj}x_n)\,.}$$

- ♣ $2n + 1 \longrightarrow H$
- ♣ $h_{ij}(x_i) \longrightarrow w_{ij}x_i$, where w_{ij} are weights
- ♣ $g_j(y) \longrightarrow v_j\sigma(y) = v_j\frac{1}{1+e^{-\alpha y}}$, where v_j are weights, and α is constant

How to use $f(\{x_i\})$?

- We should adjust weights by training NN with known Monte-Carlo events
- Traning procedure is controled by error function

$$\boxed{\chi^2 = \frac{1}{2N}\sum_{i=1}^{N}(f_i - t_i)^2}$$

- N is the number of training events
- t_i is desired NN output
- f_i is actual NN output

The less value of the error function the more precise network we have

Optimization of input variables

- Goal: maximization of information in each input i from each sample n.
  - Entropy $H(\{x_i\}) \rightarrow max$.
- Different input variables (even NNs) for different signal-background pairs
  - Use physical sense to code difference between $S \leftrightarrow B_1$,. . . , $S \leftrightarrow B_n$.

♠ Correlation of input variables

$$H(x_i, x_j) \leq H(x_i) + H(x_j)$$

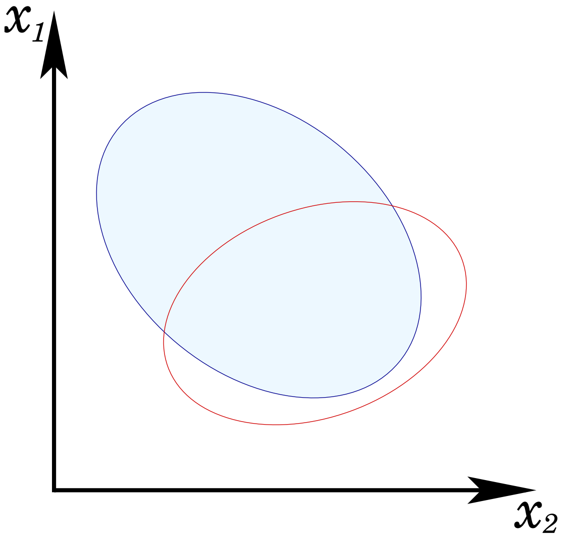
$H(x_i, x_j) = H(x_i) + H(x_j)$ if x_i and x_j are independent

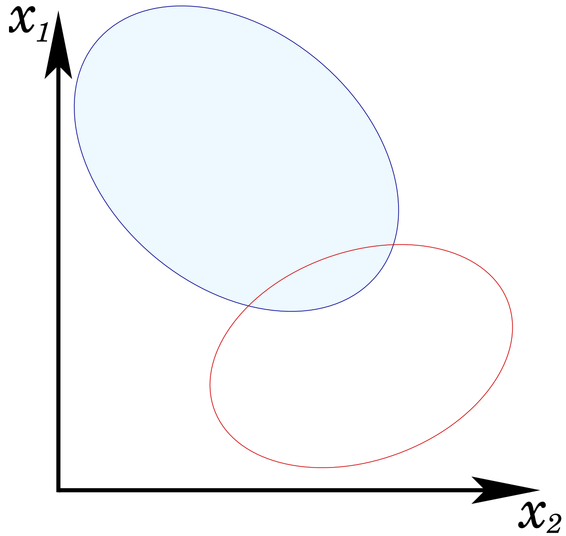
covariance matrix:

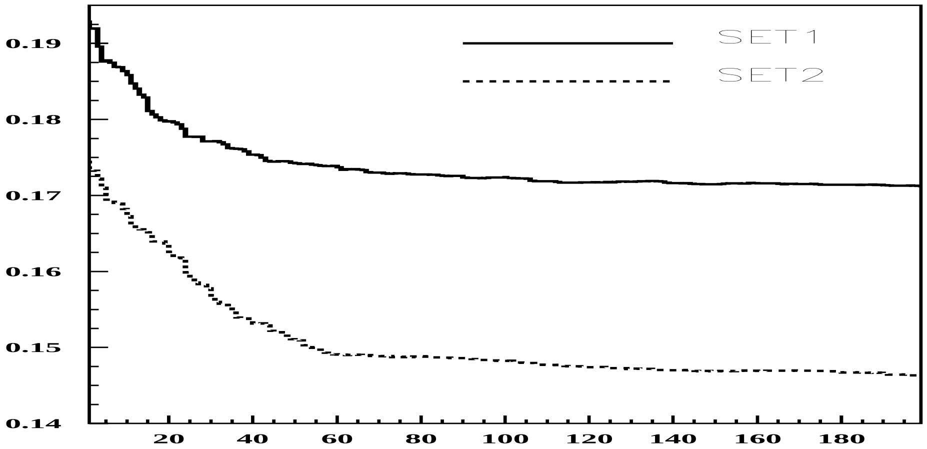
$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_i^{(k)} - \bar{x}_i^{(k)})(x_j^{(k)} - \bar{x}_j^{(k)}) \quad \forall i \neq j, \quad \bar{x}_i \equiv \frac{1}{P} \sum_{k=1}^{P} x_i$$

$$C = \left(\begin{array}{cc} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{array} \right), \qquad i, j = 1, 2$$

- Result of variable optimization:







Estimations for number of hidden units H

- Structure
  - H is number of hidden units
  - n and t are numbers of inputs and outputs
  - $W \sim Hn$ is number of weights
  - P is number of training samples

Too small H can lead to undertraining

Too large H can lead to overtraining

♠ Theoretical estimation

♣

$$\epsilon \sim \epsilon_{approx} + \epsilon_{complexity}$$

$$\epsilon_{approx} \sim 1/H \sim d/W \,, \quad \epsilon_{complexity} \sim W/P$$

$$\epsilon \sim d/W + W/P \to \min\,, \quad \text{when} \quad W \sim \sqrt{Pn}, \quad \text{i.e.} \quad \boxed{H \sim \sqrt{\frac{P}{n}}}$$

♣ $H \sim 2n + 1$ from Kolmogorov theorem

♠ Empiric estimations

$$H = 0.5(n + t) + \sqrt{P}, \qquad H = 2\sqrt{tn}$$

Conclusion

- ♠ NN can give better result if it is used correctly
- ♠ References:
  - ♣ Internet. Ask search engines.
  - ♣ smirnov@unm.edu